

Distributed Multi Class SVM for Large Data Sets

Aruna Govada
BITS-Pilani, KK Birla Goa
Campus
Goa
India
garuna@goa.bits-
pilani.ac.in

Bhavul Gauri
BITS-Pilani, KK Birla Goa
Campus
Goa
India
bhavul93@gmail.com

S.K.Sahay
BITS-Pilani, KK Birla Goa
Campus
Goa
India
ssahay@goa.bits-
pilani.ac.in

ABSTRACT

Data mining algorithms are originally designed by assuming the data is available at one centralized site. These algorithms also assume that the whole data is fit into main memory while running the algorithm. But in today's scenario the data has to be handled is distributed even geographically. Bringing the data into a centralized site is a bottleneck in terms of the bandwidth when compared with the size of the data. In this paper for multiclass SVM we propose an algorithm which builds a global SVM model by merging the local SVMs using a distributed approach(DSVM). And the global SVM will be communicated to each site and made it available for further classification. The experimental analysis has shown promising results with better accuracy when compared with both the centralized and ensemble method. The time complexity is also reduced drastically because of the parallel construction of local SVMs. The experiments are conducted by considering the data sets of size 100s to hundred of 100s which also addresses the issue of scalability.

Categories and Subject Descriptors

I.5 [Computing Methodologies]: Pattern Recognition

General Terms

Learning Model, Hyperplane

Keywords

Distributed Data Mining, MultiClass SVM, One-Vs-One(OVO)

1. INTRODUCTION

Data mining algorithms demand the data to be available at one centralized site. In today's era of massive data sets which are distributed geographically, bringing this whole data to a centralized site is almost impossible due to the limited bandwidth when compared with the size of the data. And also solving a large problem at a central site is not feasible in terms of the computational complexity.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

WCI '15, August 10 - 13, 2015, Kochi, India

©2015 ACM. ISBN 978-1-4503-3361-0/15/08\$15.00

DOI: <http://dx.doi.org/10.1145/2791405.2791534>

All traditional data mining algorithms assume that the data should fit into main memory which is a challenge for data mining algorithms in terms of scalability. [14]

In many domains like financial[15], health care[1], astronomy[11] the data is overflowing resulting in data avalanche due to the advances in data collection methodologies. In all these applications, the data may not reside at a centralized location. For example, the different sky survey telescopes which are geographically distributed must be collecting the data of common interest continuously. Mining from these massive data can not be achieved unless the data mining algorithms are capable of handling the decentralized data.[8] [9].

Distributed Data Mining (DDM) can be one of the solution for the above said problem. DDM can be achieved in two ways, Data Mining on distributed data or Distributing the Data Mining on the centralized data. In this paper we discuss the first scenario in which the data is distributed at different sites. DDM is possible on horizontal partition as well as vertical partition also. In Horizontal partition the number of attributes are constant at all n different locations but the number of instances may vary. Whereas in vertical partition the number of instances are constant at all n different locations but number of attributes may vary. [4]

In this paper the data is partitioned in horizontal manner. The proposed distributed approach is compared with the centralized method by bringing the distributed data to one central site. The multiclass SVM is achieved using One-Versus-One(OVO) approach both in centralized and distributed approach. The experimental analysis shows how our distributed approach is better than the normal approach in terms of accuracy, training time and testing time. Experiments are conducted by considering different data sets of different size. Our proposed approach could succeed in building the global SVM in case of large data sets whereas the centralized approach could not handle the data to build the training model.

The rest of the paper is organized as follows. In next section the related work is discussed. Section 3 briefly describe the binary SVM, OVO multiclass SVM. In section 4 we present our distributed approach for scalable distributed data mining to construct the merged model. In section 5 we present the experimental analysis of our algorithm. At the end in section 6 the conclusions of the paper are mentioned.

2. RELATED WORK

Though a decent amount of work has been done in multiclass SVM and parallel/distributed binary SVM individually, the research of distributed multiclass SVM needs more exploration by the research community. There is a continuous attention on SVM because it was proved as the best method in several applications even though it is computationally expensive [12].

In 1998 Han.X et al. discussed the model of coupling the estimation of class probabilities for each pair of classes [6]. The used classifiers include linear discriminant, nearest neighbors and SVM.

In 2005 Hian et al. discussed the association between the symptoms and the treatment of a patient [10]. their discussion also includes the significance of handling the huge amount with the help of data mining.

In 2010 Stefano et al. discussed the construction of a SVMs based on the Minimal Enclosing Ball(MEB) when the training data is partitioned at several locations [13]. It is shown how the union of local core-sets provides a close approximation to a global core-set from which the SVM can be recovered.

In 2011 Ahmed et al. designed a hybrid ensemble model for credit risk which combines both clustering and classification [5]. In this SVM classifiers are the members in the ensemble model.

A multiclass classification approach for large data sets is discussed by using SVM, enclosing ball(MEB) method [3]. Solving a single optimization problem for the multiclass is very expensive in terms of the time. A distributed parallel training approach is discussed for single-machine problem in [7].

In 2014 Aruna.G et al. proposed a binary tree based support vector machine [2] which reduces the number of binary classifiers when compared with OVO and OVA approaches. This algorithm is implemented in a distributed manner under HADOOP framework.

3. PRELIMINARIES

3.1 Binary SVM

Given some training data D , a set of k points of the form $D = \{(p_i, q_i) \mid p_i \in R^p, q_i \in \{-1, 1\}\}$, $i=1$ to k . The goal is to find the maximum-margin hyperplane that divides the points having $q_i = 1$ from those having $q_i = -1$.

The function of learning in binary SVM can be represented as follows. [14]

$$\min_w = \frac{\|w\|^2}{2}$$

subject to $q_i(w \cdot p_i + b) \geq 1$, $i = 1, 2, \dots, k$ where w and b are the parameters of the model for total k number of instances.

Using Lagrange multiplier method the following equation has to be solved,

$$L_p = \frac{\|w\|^2}{2} - \sum_{i=1 \dots k} \lambda_i (q_i(w \cdot p_i + b) - 1)$$

The dual version of the above problem is

$$L_D = \sum_{i=1 \dots k} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j q_i q_j p_i \cdot p_j$$

subject to

$$\lambda_i \geq 0$$

$$\lambda_i (q_i(w \cdot p_i + b) - 1) = 0$$

where λ_i are known as the Lagrange multipliers.

By solving this dual problem, SVM(hyperplane) will be found. Once the training model is built, the class label of a testing object z can be predicted as follows.

$$f(z) = \text{sign} \sum_{i=1 \dots k} (\lambda_i q_i p_i \cdot z + b)$$

if $f(z) \geq 0$ z will be predicted as +ve class else -ve class.

3.2 One-Versus-One(OVO):A MultiClass SVM

Multiclass classification is the problem of classifying instances into one class label among the N -class labels. Build $N(N-1)/2$ classifiers one classifier to distinguish each pair of classes i and j . Let f_{ij} be the classifier where class i were +ve examples and class j were -ve. Classify using

$$f(x) = \arg \max_i \left(\sum_j f_{ij}(x) \right)$$

This way in OVO, each class is compared to every other class. A binary classifier is built to differentiate between each pair of classes, while discarding the rest of the classes. When an unseen object has to be classified into one of the class, a voting is made among the classifiers and the class with the maximum number of votes will be considered as the best choice.

4. THE PROPOSED APPROACH

The data be distributed among n sites with equal number of attributes but varied in number of instances.

1. Let the data is distributed among n sites .

$$[\mathbf{X}]_{p \times q} = (X_1, X_2, X_3, \dots, X_n)$$

where data X_j is a $p_j \times q$ matrix residing at the site S_j and $p = \sum_{j=1}^n p_j$

2. Build the local SVM models $SVM_1, SVM_2, \dots, SVM_n$ at all n sites individually.

3. Construct a directed graph as follows.

- Each site is a vertex .
- Edge (i->j) refers to the training model SVM_i w.r.t test data at site j .(where $i= 1$ to n , $j=1$ to n , $i \neq j$)
- Label the edge with Accuracy of (SVM_i ,j) .

// Merging the local models into a global model

4. For each vertex j, Find out the Maximum labeled edge among all the edges from i to j, where $i= 1$ to n , $i \neq j$. and store the values in $n \times 2$ matrix as follows.

- For (k= 1 to n)
 - Best [k][1]=i;
 - Best [k][2]=j; // Decides the best model SVM_i w.r.t test data at site j

5. Find out the element which is having the maximum frequency among Best[i][1] , where $i=1$ to n . And the corresponding SVM model is decided as the global/merged model.

The architecture of the proposed approach is shown in figure 1, where the global model is built by merging the local SVMs. And the global model made available at each local site by transmitting it so that it can be used in future for classifying unseen objects.

4.1 Graphical Representation

The training data at n sites can be constructed as a graph and shown in Figure 2. Each site is considered as a vertex and the edge from the vertex i to vertex j represents the accuracy of the SVM model of the training data at site i w.r.t the test data at site j .

The *Accuracy Matrix* A_{ij} is an $n \times n$ matrix which can be written as follows.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & \dots & a_{nn} \end{pmatrix}$$

To get the final global SVM find out the Maximum of each column and note down the corresponding i value of SVM. Among these n-Max values (n-SVMs), choose the SVM which will get the maximum count as the final global model.

5. EXPERIMENTAL SECTION

We implemented the algorithm on three data sets Mfeat-Fac, Pendigits and Sloan Digital Sky Survey (SDSS). Mfeat and Pendigit are taken from UCI machine learning repository <https://archive.ics.uci.edu/ml/datasets.html>. SDSS is an astronomical catalog taken from <http://skyserver.sdss.org/dr7/en/tools/search/sql.asp> based on different conditions of its attributes. Mfeat, Pendigit and SDSS data sets are divided into 3,4,4 partitions respectively and implemented DSVM.

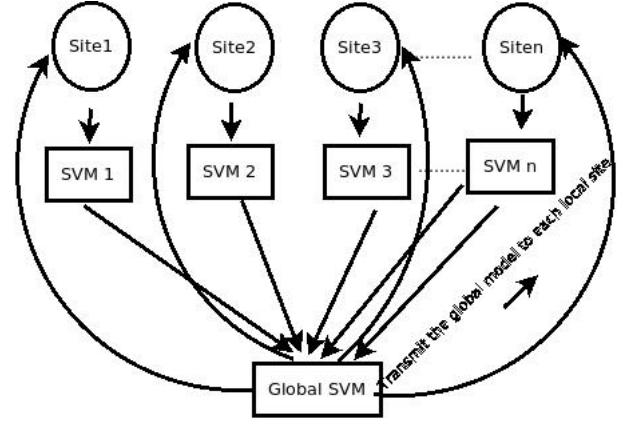


Figure 1: The Architecture

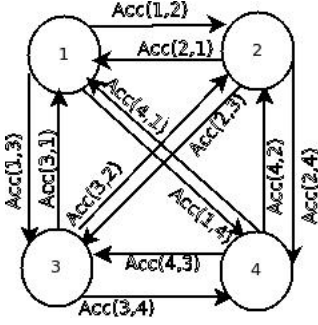


Figure 2: SVMs of i^{th} site w.r.t test data of j^{th} site.

In our analysis the DSVM is compared with centralized SVM and Ensemble SVM. The results show the better accuracy with reduced training time and testing time.

In table 1 the description of the data sets that are considered for analysis is given. In table 2 the training accuracy and the training time is given. The error of global model will not exceed the error of Max(Local Models) as the data is independent where as in ensemble model this is not guaranteed because the same samples can be repeated at different sites. In table 3, 4, 5 the construction of the global models of Mfeat-Fac, Pendigits, SDSS data sets are given respectively. If we observe ,

The accuracy matrix of Mfeat-Fac with three sites

$$\begin{pmatrix} -1 & 96 & 96.8 \\ 98.2 & -1 & 98 \\ 97 & 96.3 & -1 \end{pmatrix}$$

The best model will be SVM_2 as it has the maximum count and is shown in table 3.

The accuracy matrix of Pendigits with four sites

$$\begin{pmatrix} -1 & 99.45 & 99.70 & 99.30 \\ 99.50 & -1 & 99.60 & 99.40 \\ 99.60 & 99.25 & -1 & 99.20 \\ 99.50 & 99.35 & 99.5 & -1 \end{pmatrix}$$

The best model will be SVM_1 as it has the maximum count and is shown in table 4.

Table 3: The Best Global Model of Mfeat-Fac: SVM_2

Test Data	Training Model	Accuracy	Best Model
Site ₁	SVM_2	98.2	SVM_2
	SVM_3	97.0	
Site ₂	SVM_1	96.0	SVM_3
	SVM_3	96.3	
Site ₃	SVM_1	96.8	SVM_2
	SVM_2	98.0	

Table 4: The Best Global Model of Pendigits: SVM_1

Test Data	Training Model	Accuracy	Best Model
Site ₁	SVM_2	99.50	SVM_3
	SVM_3	99.60	
	SVM_4	99.50	
Site ₂	SVM_1	99.45	SVM_1
	SVM_3	99.25	
	SVM_4	99.35	
Site ₃	SVM_1	99.70	SVM_1
	SVM_2	99.60	
	SVM_4	99.50	
Site ₄	SVM_1	99.30	SVM_2
	SVM_2	99.40	
	SVM_3	99.20	

The accuracy matrix of SDSS with four sites

$$\begin{pmatrix} -1 & 68.38 & 68.39 & 68.52 \\ 37.45 & -1 & 89.16 & 89.02 \\ 37.45 & 89.53 & -1 & 89.37 \\ 37.47 & 89.63 & 89.66 & -1 \end{pmatrix}$$

The best model will be SVM_4 as it has the maximum count and is shown in table 5.

In table 6 the ensemble method is computed for all 3 data sets. In ensemble method the test data has to be tested every time with all the available training models and the class label will be decided by the voting approach. Where as in DSVM, the final global model is merged from all local models. And whenever an unseen object has to be classified, it will be tested with only one global model. Hence the testing time of DSVM is reduced when compared with Ensemble SVM.

Finally in Table 7, DSVM is compared with OVO multi-class SVM with centralized and ensemble model. The results show that the accuracy of DSVM is equivalent to centralized SVM with reduced training time and testing time. The training time of DSVM is considered as the time of Max(local SVMs) as the local SVMs can be constructed in a parallel manner. For the data set SDSS the system crashed during the training of SVM for centralized method but DSVM built the training model without any hurdle, hence it is scalable.

6. CONCLUSIONS

We propose an algorithm DSVM which builds a global SVM by merging the local SVMs that are distributed at different sites. Experimental results show that the performance of DSVM is better than the centralized and Ensemble model both in accuracy and training, testing time. DSVM is also capable of handling large data sets, hence scalable. Though

Table 5: The Best Global Model of SDSS : SVM_4

Test Data	Training Model	Accuracy	Best Model
Site ₁	SVM_2	37.45	SVM_4
	SVM_3	37.45	
	SVM_4	37.47	
Site ₂	SVM_1	68.38	SVM_4
	SVM_3	89.53	
	SVM_4	89.63	
Site ₃	SVM_1	68.39	SVM_4
	SVM_2	89.16	
	SVM_4	89.66	
Site ₄	SVM_1	68.52	SVM_3
	SVM_2	89.02	
	SVM_3	89.37	

Table 6: The Ensemble Model of Data Sets

Data Set	Local Site	Accuracy	Voting Model
Mfeat-Fac	Site ₁	97.50	SVM_2
	Site ₂	98.00	
	Site ₃	97.00	
Pendigits	Site ₁	10.40	SVM_1
	Site ₂	10.37	
	Site ₃	10.37	
	Site ₄	10.37	
SDSS	Site ₁	85.80	SVM_4
	Site ₂	53.30	
	Site ₃	53.30	
	Site ₄	87.10	

Ensemble method also can handle large data sets at the time of training, testing time will be very costly as it has to be tested with every training model (which are available at different locations) and follow voting mechanism. But DSVM will have only one global model and it is scalable for training as well as for testing also. Further enhancement can be done by considering the vertical partition of the data at different sites.

7. ACKNOWLEDGMENTS

We are thankful for the support provided by the Department of Computer Science and Informations Systems, BITS, Pilani, K.K. Birla Goa Campus to carry out the experimental analysis.

8. REFERENCES

- [1] C. R. Y. K. al. Prediction of conversion from mild cognitive impairment to alzheimer disease based on bayesian data mining with ensemble learning. *The Neuroradiology Journal*, 25(1), 2012.
- [2] G. Aruna, Ranjani, Aditi, and S. Sahay. A novel approach to distributed mutli-class svm. *Transactions on Machine Learning and Artificial Intelligence*, 2(5):72–79, October 2014.
- [3] J. Cervantes, X. Li, and W. Yu. Multi-class svm for large data sets considering models of classes distribution. In *International Conference on Data Mining*, pages 257–268. DMIN, July 2008.
- [4] H. Dutta and et al. Distributed top-k outlier detection from astronomy catalogs using the demac system. In *SDM Proceedings*, pages 473–476. SIAM, 2007.

Table 1: The description of the data sets used.

Dataset	Features	Training Sizes	Testing Size	Class Labels	At site1	At site 2	At site 3	At site 4
Mfeat-Fac	216	1800	200	10	500	800	500	-
Pendigits	16	7514	3478	9	1800	2000	1494	2200
SDSS	5	175000	1000	5	20000	30000	50000	75000

Table 2: Training Accuracy and Training Time of Data sets at Distributed Sites.

Dataset	Site 1		Site 2		Site 3		Site 4	
	Accuracy	Time	Accuracy	Time	Accuracy	Time	Accuracy	Time
Mfeat-Fac	100	0.1365	100	0.256	100	0.143	-	-
Pendigits	100	0.062	99.53	0.110	100	0.049	100	1.0132
SDSS	100	49.996	89.09	115.76	89.13	369.93	89.76	1728.097

Table 7: The Comparison Of the DSVM with Centralized and Ensemble Models

Dataset	Centralized			Ensemble			DSVM		
	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time	Te.Time	Accuracy	Tr.Time	Te.Time
Mfeat-Fac	99	0.667	0.133	98	0.234	0.204	98	0.234	0.077
Pendigits	10.40	0.477	0.190	10.40	0.243	0.868	10.40	0.243	0.135
SDSS	-	-	-	89	227.24	5.52	89	227.24	2.94

- [5] A. Ghodselahi. A hybrid support vector machine ensemble model for credit scoring. *International Journal of Computer Applications*, 17(5):975–8887, March 2011.
- [6] B. Han.X. Classification by pairwise coupling. *Advances in Neural Information Processing*, 10(2):291–301, June 1998.
- [7] B. Han.X. Dcmsvm: Distributed parallel training for single-machine multiclass classifiers. In *Computer Vision and Pattern recognition Proceedings*, pages 3554–3561. IEEE, June 2012.
- [8] D. K. K. Bhaduri, and et al. Scalable distributed change detection from astronomy data streams using local, asynchronous eigen monitoring algorithms. In *SDM Proceedings*, pages 247–258. SIAM, 2009.
- [9] H. Kargupta, C. Gianella, and K. Sivakumar. Distributed data mining for earth and space science applications. In *Proceeding of the fourth annual Earth Science Technology conference (ESTC)*, June 2004.
- [10] H. C. Kob and G. Tan. Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2):64–72.
- [11] R. Mallik, N. Sarda, and H. Kargupta. Distributed data mining for sustainable smart grids. In *Proceedings of the Sustainable KDD Workshop. KDD 2011*. ACM, 2011.
- [12] J. D. M. Rennie and R. Rifkin. Improving multiclass text classification with the support vector machine. Technical report, Massachusetts Institute of Technology, 2001.
- [13] C. S. Stefano Lodi, Ricardo Nanculef. Single-pass distributed learning of multi-class svms using core-sets. In *SDM Proceedings*, pages 257–268. SIAM, 2010.
- [14] P.-N. Tan, V. Kumar, and M. Steinbach. *Introduction to Data Mining*. Pearson, NewDelhi,India, 2012.
- [15] J. Wang and H. Wang. Application of data mining in the financial data forecasting. *Advanced Intelligent*

Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, Lecture Notes in Computer Science, 5226(1):954–961, 2008.